Sun's paper targets Chinese overlapping ambiguity resolution. They use a method based on manually written rules and probabilities with supervised training.

ISSN 1000-0054 清华大学学报（自然科学版）1999 年 第 39 卷 第 5 期 27/34
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 1999, Vol. 39, No. 5 101～103

# 消解中文三字长交集型分词歧义的算法*

孙茂松， 左正平， 黄昌宁

清华大学 计算机科学与技术系；智能技术与系统国家重点实验室，北京 100084

**文 摘** 汉语自动分词在中文信息处理现实应用中占据着十分重要的位置。三字长交集型分词歧义是分词歧义的主要类型之一，在真实文本中的出现频率相当高。提出了一种针对这种分词歧义的消解算法，回避了训练代价比较高昂的词性信息而仅仅利用了词的概率信息及某些具有特定性质的常用字集合。从一个 60 万字的汉语语料库中抽取出全部不同的三字长交集型分词歧义共 5367 个作为测试样本。实验结果表明，该算法的消解正确率达到了 92.07%，基本可以满足实用型中文信息处理系统的需要。

**关键词** 计算语言学；中文信息处理；汉语自动分词；交集型分词歧义；分词歧义消解算法

**分类号** TP 391

汉语自动分词在自然语言处理应用系统中(如中文文本检索、汉字及语音识别、文语转换等)占据着十分重要的位置[1]。虽然国内外的相关研究已有不少，但真正能达到实用要求的系统严格说来还没有[2]。困难在于：自然语言十分复杂，极富变化，很难纳入一个统一整齐的计算模型中。可能的解决之道是：必须逐个分解出其中蕴涵的子问题，进行深入、细致、有针对性的研究。

交集型分词歧义是自动分词系统遇到的主要歧义类型。而三字长交集型分词歧义又是交集型分词歧义的主要类型之一。根据我们对一个 1 亿字汉语语料库的观察，三字长交集型分词歧义就静态个数而言占全部交集型分词歧义的 33.29%，动态覆盖率更占全部交集型分词歧义的 49.76%。本文提出了一种解决简单有效的办法。能够满足实用系统的需要。我们已把它作为一个功能模块嵌入我们设计的面向非受限文本的汉语自动分词系统中[3]。

收稿日期：1998-10-18

第一作者：男，1962 年生，副教授

* 基金项目：国家自然科学基金重点项目（69433010）

## 1 词概率及词性信息在解决交集型分词歧义中的作用

**定义** 给定一部汉语词典 $D$ 及任意汉字字串 $S'_{3x}=ABC$ ($A,B,C$ 为汉字)，如果满足 $AB\subset D$ 且 $BC\subset D$，则称 $S'_{3x}$ 为三字长交集型分词歧义，记作 $S_{3x}$。

从形式上看，$S_{3x}$ 有两条可能的切分路径 $AB/C$ 及 $A/BC$。歧义的解决通常要在比 $S_{3x}$ 更大的语言环境—上下文乃至句子 $S$ 中考量。

设 $W=w_1w_2\cdots w_m$ 是 $S$ 的可能切分结果之一(对应一条从起点到终点的词路径)，$T=t_1\cdots t_m$ 为词 $w_1\cdots w_m$ 的词性标记串(对应一条词性标记路径)，通常利用 $P(W,T)$ 对 $W$ 的似然性进行评价，并且认为具有最大 $P(W,T)$ 值的 $W$ 为 $S$ 的正确切分。$S_{3x}$ 的切分自然随之而定。

将 $S$ 视作一阶 Markov 链，则 $P(W,T)$ 可用下式评价：

方法 1 词概率+词性 Bigram 法

$$P(W,T)=P(T)P(W|T)\approx\prod_{i=1}^{m}P(t_i|t_{i-1})P(w_i|t_i)=\prod_{i=1}^{m}P(t_i|t_{i-1})\frac{P(t_i|w_i)P(w_i)}{P(t_i)}=\prod_{i=1}^{m}\frac{P(t_i|t_{i-1})P(t_i|w_i)P(w_i)}{P(t_i)}$$

为进行比较，依次忽略该式中词性和词概率信息，于是得到两个简化的评价公式：

方法 2 词概率法

$$P(W,T)=\prod_{i=1}^{m}P(w_i)$$

方法 3 词性 Bigram 法

$$P(W,T)=\prod_{i=1}^{m}\frac{P(t_i|t_{i-1})P(t_i|w_i)}{P(t_i)}$$

显然,方法 1、方法 3 与上下文有关,方法 2 则退化为零阶 Markov 链,与上下文无关。

对每一种方法均可定义三个有效性指标。

令 $n_1$ 表示待处理文本中的全部分词歧义个数,$n_2$ 表示某种方法能够评价的分词歧义个数,$n_3$ 表示该方法评价正确的分词歧义个数,则方法的适用率 $\eta_1$、适用正确率 $\eta_2$、正确率 $\eta_3$ 分别为

$$\eta_1 = \frac{n_2}{n_1}, \quad \eta_2 = \frac{n_3}{n_2}, \quad \eta_3 = \frac{n_3}{n_1}.$$

利用了一个 60 万字左右的汉语熟语料库(经过人工分词及词性标注处理),训练方法 1、方法 2、方法 3 所需的各项统计参数。另外,还从该语料库中抽取出全部不同的 $S_{3x}$ 共 5367 个,从而形成一个集合 $\{S^0_{3x}\}$ 作为本文实验的测试对象。对测试集 $\{S^0_{3x}\}$ 分别运用这三种方法,得到表 1 所示结果(注:如果对两条可能的切分路径,根据某种方法给出的评价值相等,则认为此时该方法不适用)。

**表 1 3 种方法性能比较**

| 方法 | $100\times\eta_1$ | $100\times\eta_2$ | $100\times\eta_3$ |
|---|---|---|---|
| 词概率法 | 100.00 | 85.46 | 85.46 |
| 词性 Bigram 法 | 97.11 | 71.37 | 69.31 |
| 词概率+词性 Bigram 法 | 100.00 | 86.94 | 86.94 |

可分 4 种情形:

情形 1:词概率法评价正确、词性 Bigram 法评价也正确的有 3594 个(此时词概率+词性 Bigram 法的评价一定正确);

情形 2:词概率法评价正确、词性 Bigram 法评价错误的有 983 个(其中词概率+词性 Bigram 法评价正确的有 915 个,评价错误的有 68 个);

情形 3:词概率法评价错误、词性 Bigram 法评价正确的有 232 个(其中词概率+词性 Bigram 法评价正确的有 158 个,评价错误的有 74 个);

情形 4:词概率法评价错误、词性 Bigram 法评价也错误的有 558 个(此时词概率+词性 Bigram 法的评价一定错误)。

从此结果来看,简单的词概率法与词概率+词性 Bigram 法的正确率相差不大,仅 1.48%。由于在词概率法中加入词性 Bigram 后仅能解决情形 3 中的分词歧义,而这部分分词歧义占总的分词歧义的比例本来就较小(232/5 367 = 4.32%),因此词性 Bigram 法对词概率法的纠错能力有限(况且词性 Bigram 法还造成情形 2 中部分词概率法原可正确处理的分词歧义产生错误)。

另外,由于建立获取词概率信息所需要的分词熟语料库相对容易(退一步讲,假如只有未经分词的生语料库,还可以直接用 $w_i$ 在语料库中的字串频来近似求得 $P(w_i)$,根据我们的经验,此时切分正确率仅略为下降),但是引入词性信息后,需要知道词性的条件概率 $P(t_i|t_{i-1})$ 和每个词的词性分布概率 $P(t_i|w_i)$,而这必须通过一个大规模的、预经人工分词且词性标注的熟语料库才能加以训练,工作量迅猛增加,代价更加高昂。因此,我们认为,词性信息对解决分词歧义意义不大,就一般的自然语言处理应用系统而言,简单的词概率法就已基本够用。

## 2 对词概率法的进一步考察

设对应 $W=w_1w_2\cdots w_m$ 的词概率评价函数为

$$f(w_1,w_2,\cdots,w_m) = lb(\prod_{i=1}^{m} P(w_i)),$$

则词概率法可描述为:

对 $S_{3x}=ABC$,其切分取作

$$\begin{cases} AB/C, \text{如果 } f(AB,C) > f(A,BC); \\ A/BC, \text{如果 } f(AB,C) < f(A,BC); \\ \text{未定, 如果 } f(AB,C) = f(A,BC). \end{cases}$$

可以想象,$f(AB,C)$ 和 $f(A,BC)$ 之间的差值不同,词概率法对 $S_{3x}$ 的切分正确率也应是不同的。考察所有在某一个固定差值 $x$ 附近的 $S_{3x}$,即所有满足 $x-\Delta<\delta=|f(AB,C)-f(A,BC)|<x+\Delta$ 的 $S_{3x}$ 组成的集合($\Delta$ 是一个小正数),如果词概率法对该集合的切分正确率记为 $c(x)$,则实验可得 $c(x)$ 随 $\delta$ 的变化曲线如图 1。

图 1 显示,当 $\delta>2.5$ 时,词概率法的切分正确率很高,并且基本稳定;当 $\delta\leqslant 2.5$ 时,切分正确率则急剧下降。

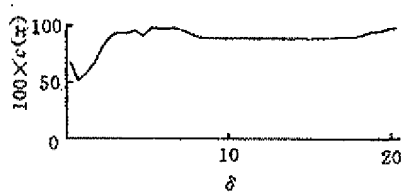


图 1 $c(x)$ 与 $\delta$ 的关系

于是,可对词概率法改进如下($t_0$ 为一阈值):

对 $S_{3x}=ABC$,其切分取作

$$\begin{cases} AB/C, \text{如果 } f(AB,C) - f(A,BC) > t_0; \\ A/BC, \text{如果 } f(AB,C) - f(A,BC) < -t_0; \\ \text{未定, 如果 } |f(AB,C) - f(A,BC)| \leqslant t_0. \end{cases}$$

改进后的词概率法,其适用正确率随 $t_0$ 增大将会提高,但同时适用率将会降低。按改进的词概率法

对测试集$\{S^0_{3x}\}$进行切分，可得适用正确率$\eta_2$随适用率$\eta_1$的变化曲线如图2。

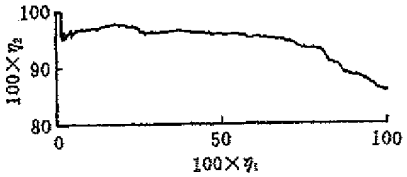


图2 改进词概率法适用正确率与适用率的关系

当$t_0$取1.92时，正确率和适用正确率获得比较满意的折衷：当适用率为67.97%时，适用正确率为95.01%，对应的正确率为64.58%.

## 3 常用字分合法

对$S_{3x}=ABC$，根据分词语料库对$\{S^0_{3x}\}$的人工切分结果，可搜集到6个关于首字$A$、中字$B$、尾字$C$的字表：字表$L_1$为切分为单字词频率很高的首字$A$的集合，字表$L_2$为切分为单字词频率很低的首字$A$的集合，字表$L_3$为与首字$A$结合成词频率很高的中字$B$的集合，字表$L_4$为与尾字$C$结合成词频率很高的中字$B$的集合，字表$L_5$为切分为单字词频率很高的尾字$C$的集合，字表$L_6$为切分为单字词频率很低的尾字$C$的集合(见表2)。

设$L$是上述任一汉字表，$Z$是任意汉字，令

$$R_L(Z)=\begin{cases}0, & Z\notin L;\\ 1, & Z\in L.\end{cases}$$

$$k_1=R_{L1}(A)+R_{L4}(B)+R_{L6}(C),$$

$$k_2=R_{L2}(A)+R_{L3}(B)+R_{L5}(C).$$

则常用字分合法可描述为

对$S_{3x}=ABC$，其切分取作

$$\begin{cases}A/BC, & \text{如果 } k_1>k_2;\\ AB/C, & \text{如果 } k_1<k_2;\\ \text{未定}, & \text{其它.}\end{cases}$$

利用常用字分合法对$\{S^0_{3x}\}$进行切分：适用率为73.44%，适用正确率为92.87%。对应的正确率为68.20%.

## 4 算法：词概率法与常用字分合法的组合

综合改进后词概率法和常用字分合法，就得到一种能够处理$S_{3x}$的简单、有效的算法。

对任意的$S_{3x}$

1) 先用改进后词概率法切分；

2) 如果1)不能处理，则用常用字分合法切分；

3) 如果1),2)均不能处理，则退回来用改进前词概率法切分。

实验表明，步骤(1)+(2)对$\{S^0_{3x}\}$的适用率扩大为88.87%，适用正确率为92.90%(其中改进后词概率法的适用率为67.97%，适用正确率为95.01%。对改进后词概率法不能处理的剩余部分，常用字分合法的适用率为65.19%，适用正确率为86.05%)，正确率为82.56%。整个算法的适用率为100%，总的正确率则达到了92.07%，同改进前的词概率法及词概率+词性Bigram法的正确率相比较，分别提高了6.61%和5.13%.

## 5 结束语

本文提出了一种针对三字长交集型分词歧义的消解算法。该算法追求的目标是比较高的"性能价格比"，并不刻意强调方法的深奥与完美，所以在设计思想上回避了训练代价相对高昂的词性信息，而仅仅利用了词的概率信息及某些具有特定性质的常用字集合。经验显示，这个解决方案虽然十分简单(就本质而言，它是不考虑上下文制约关系的零阶Markov模型)，但却有效，基本可以适应中文信息应用系统处理三字长交集型分词歧义的现实要求。

表2 关于首、中、尾字的6张特殊字表

| 字表 | 字例 | 与字表有关的分词歧义示例 |
|---|---|---|
| $L_1$ | 不在和上对有的为并新高年着从多等过了当以是与会就做更日受外无种 | 不成熟，从今年，上岗位，为人民 |
| $L_2$ | 适分通发实需转参合集节今决生特工战主表部常规简没农区身统完制公 | 发生于，分别是，战争论，主要地 |
| $L_3$ | 面定别年现路求长动目备除口立次度空入为业 | 需求是，去年底，机动力，科长处 |
| $L_4$ | 生数政工调内事气保处情经前任下表等防关可美清意公军力难电及具流其世市限相效眼优指装话配士 | 大调整，此前提，多事物，从政治 |
| $L_5$ | 的地和了是为时上在中有给性新将向会里数敌对部下带或派日小则从而分 | 表面的，沉着地，参谋和，成就是 |
| $L_6$ | 量理民动兵产业行度心予制员备国节去生务队物作持合空年求围形治实道方府格工活积家进类立律率命校展整证值置质见 | 对调动，大力量，并进行，同行业 |

(下转第107页)